ACS Combinatorial Science

Research Article

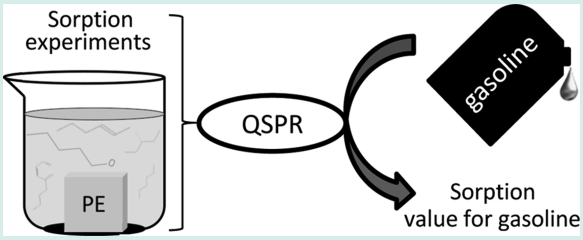
pubs.acs.org/acscombsci

Prediction of Alternative Gasoline Sorption in a Semicrystalline Poly(ethylene)

Nicolas Villanueva,[†,‡] Bruno Flaconnèche,[†] and Benoit Creton*[,†]

[†]IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France

ABSTRACT: In this work, we first report the acquisition of new experimental data and then the development of quantitative structure−property relationships on the basis of sorption values for neat compounds and up to quinary mixtures of some hydrocarbons, alcohols, and ethers, in a semicrystalline poly(ethylene). Two machine learning methods (i.e., genetic function approximation and support vector machines) and two families of descriptors (i.e., functional group counts and substructural molecular fragments) were used to derive predictive models. Models were then used to predict sorption variations when increasing the number of carbon atoms in a series of hydrocarbons and for *n*-alkan-1-ols. In addition to the performed internal/external validations, the model was further tested for surrogate gasolines containing ca. 300 compounds, and predicted sorption values were in excellent agreement with experimental data (R^2 = 0.940).



KEYWORDS: *hydrocarbon, alcohol, ether, gasoline, sorption, semicrystalline poly(ethylene), QSPR*

INTRODUCTION

The systematic use of alternative fuels and especially biofuels respecting durability criteria appears as a promising solution to the problem of climate change, reduction of greenhouse gas emissions and wastes, fuel availability at a reasonable cost, etc.[1] Biofuels can be considered as mixtures of renewable molecules, such as normal and iso-paraffins, naphthenic and aromatic compounds, normal and iso-olefins, alcohols, ethers, and esters.[2] The composition of fuel blends constantly evolves and also varies from one country to another. Materials compatibility is of major concern especially as the fuel composition changes and with the consideration of oxygenated compounds in the pool of renewable molecules. Indeed, the introduction of this latter family of chemicals may lead to problems of corrosion of metallic materials[3] and degradation of polymeric materials.[4] The use of corrosion inhibitors as fuel additives has been proposed to reduce corrosiveness effects on metallic pieces in contact with alternative fuels.[5]

Polymeric materials in contact with biofuels may be subject to deformations such as swelling caused by solvent ingress in their structure (permeation) leading to strong modifications and loss of their initial physical and chemical properties. One of the proposed solutions to address this problem is the use of a multilayer structure including barrier polymers.[6] Poly(ethylene) and poly(amide) are typically polymeric materials encoutered for tank and fuel line applications, respectively. A limited number of works have been published in the literature on the area of compatibility of such polymers with alternative fuels, and experimental data are scarce.[4,7−11]

The permeability (P), which is defined in eq 1 as the product of the diffusion (D) and the solubility (S), is an indicator of the amount of solvent ingressed in tested polymeric materials.

$$P = DS \tag{1}$$

Maru et al. compared interactions between biodiesels and two kinds of materials used in storage and automotive tanks showing that properties of high density poly(ethylene) (HDPE) are more affected by biodiesel compared with those of carbon steel.[4] Berlanga-Labari et al. studied the compatibility of HDPE with gasoline blends containing 5% and 10% ethanol using infrared spectroscopy and sorption tests.[7] The authors reported only small variations of HDPE physical and chemical properties after immersion tests of thousands of hours duration. During a screening procedure, the systematic use of such experiments to test the compatibility of new polymeric materials with new alternative fuels seems to be unrealistic, and the development and use of robust predictive models should be more appropriate.

Works in the literature deal with the quantitative prediction of permeability for polymeric materials.[12−17] Memari et al. have studied gas mixture (H_2, CO_2, and CH_4) sorptions in poly(ethylene) below its melting point using Monte Carlo (MC) molecular simulations.[14,15,17] From a technical point of view, the consideration of larger permeant molecules such as hydrocarbons, alcohols, or ethers may lead to problems of insertions and thus problems of convergence. Moreover, the inclusion of molecular simulation techniques such as MC or molecular dynamics (MD) into high throughput screening procedures still represents a challenge, and approaches based on the concept of quantitative structure−property relationship (QSPR) appear more appropriate.[18] On the basis of eq 1, Teplyakov and Mears proposed the following empirical relation to predict the permeability of inorganic gases such as N_2, O_2,

Received: June 15, 2015
Revised: September 3, 2015
Published: September 8, 2015

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

CO_2, and C_1-C_4 hydrocarbons gases through different polymers:[19]

$$\log P = \log D + \log S = K_1 - K_2 d_{ef}^{\ 2} + K_3 - K_4(\epsilon/k) \quad (2)$$

where d_{ef} is defined as the effective molecular diameter determined by comparing log D data for various gas−polymer couples, ϵ/k values were determined from log S data for specific polymers, and K_1, K_2, and K_3, and K_4 are coefficients regressed on log D and log S data, respectively.[19] However, it is necessary to explicitly understand the relation between both polymer and permeant molecular structures for the development of suitable membrane materials. Hence, Patil et al. showed that it is possible to correlate the permeability of gases or liquids through polymers with some molecular features such as molecular connectivity, molecular polarizability, and molecular weight.[20] These authors proposed three multilinear models each applicable in the cases of (i) a specific polymer (poly(vinyltrimethylsilane), poly(isoprene), and poly(urethane)) and (ii) solely for neat hydrocarbons and alcohols. The group of Izák et al. have during the last 10 years continuously experimentally and theoretically investigated gas and liquid sorption into polymers.[21−25] Very recently, Randová et al. have proposed a method based on thermodynamic aspects to predict sorption of pure organic liquids into various polymers.[26] These last years, we have devoted great effort to the development of QSPR based models for the prediction of various property values and shown that such approaches are applicable to simple mixtures and relevant for alternative fuel blends.[1,2,27−30]

In the present work, we report the acquisition of new experimental sorption values for neat compounds and up to quinary mixtures of hydrocarbons, alcohols, and ethers in a semicrystalline poly(ethylene). Additionaly, we present QSPR based models developed using two machine learning methods, two different families of descriptors, and the new experimental data. The so obtained models are then used to predict the sorption of some alternative fuels in poly(ethylene). The paper is organized as follows: we present experimental data measurements and the strategy followed to build new QSPR based models, the predictive performance of models is then exposed and discussed, and the last section gives our conclusions.

MATERIALS AND METHODS

Experimental Procedure. *Materials and Sample Preparation.* A medium density poly(ethylene) (MDPE) was used during our measurements. MDPE pieces used were in the form of plane sheets with thickness of ca. 3 mm, extracted from an extruded band. This MDPE presents excellent resistances to crack formation and growth and incorporates an optimized formulation of additives (antioxidants) to provide a long-term stability in service. Its density (ρ = 0.937 ± 0.02 g·cm^{-3}) was determined by weighing samples both in water and in toluene. The crystalline fraction of the MDPE was obtained from differential scanning calorimetric measurements with a heating rate of 10 °C min^{-1} and was found to be 50.2% ± 0.5%.

All pure liquids (linear and branched alkanes, naphthenic and aromatic hydrocarbons, alcohols and ethers) with high purity grade were purchased from Merck, and no additional purification was performed. All liquids were used to generate from binary to quinary mixtures for which we measured sorption values in poly(ethylene). For fluids expected with high sorption values in poly(ethylene) (i.e., normal paraffins, isoparaffins, olefins, and naphthenic and aromatic compounds), we have used small rectangular pieces (40 mm × 8 mm × 3 mm), with a mass of about 0.9 g. For fluids expected to be poorly soluble in poly(ethylene) (i.e., oxygenated fluids), we have used larger and heavier (mass of about 6 g) rectangular samples to increase the precision on measurements.

Three commercial gasolines were used as bases to elaborate 12 blends with various amounts of ethanol, methyl *tert*-butyl ether (MTBE), and ethyl *tert*-butyl ether (ETBE). The Carburane software, developed at IFP Energies nouvelles,[31] was used to post-treat gas chromatographic data of gasolines in order to identify and quantify their representative components.[32,33] This approach allows us to identify precisely a representative composition for a gasoline (a mixture containing ca. 300 compounds) which can be used for instance, as input of atomic/molecular-level simulations[34] or QSPR models. Compositions of the three gasolines are illustrated with circle charts on Figure 1, showing, for instance, the high paraffin content

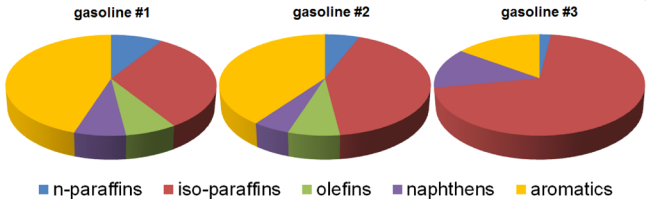


Figure 1. Circle charts of compositions for the three used commercial gasolines.

(ca. 75%) of one gasoline compared with the two remaining ones (ca. 45%). The highly paraffinic gasoline contains low amounts of unsaturated compounds compared with the two other considered gasolines.

Sorption Measurements. Measurements of liquid sorption in a MDPE were performed using a gravimetric method. The polymer samples (pieces of 0.9 or 6 g, as discussed previously) were immersed in a large excess of studied liquid in a closed glass vessel. Glass vessels were placed at ambient temperature (20 ± 1 °C) in an air-conditioned laboratory, for the duration of the sorption experiments. Polymer samples were daily removed from liquid, wiped with precaution, and weighed. The necessary time for polymer−liquid system equilibration depends on the studied system. The shortest equilibration time was observed for the toluene for which a total duration of approximately 200 h was measured. In contrast, the longest equilibration time observed was measured after up to 2000 h of immersion for the 2,2,4-trimethylpentane. Figure 2 shows three sorption curves for toluene, *n*-hexane, and 2,2,4-trimethylpentane. Figure 2 demonstrates the excellent repeatability of our measurements. Indeed, some selected tests were doubled to obtain an estimate of the repeatability of the sorption experiments. For instance, we obtained coefficients of variation of repeatability of 1% in the cases of toluene and *n*-hexane and 1.5% in the case of 2,2,4-trimethylpentane.

Modeling Methods. *Data Sets.* One of the keystones for the accuracy of predictive QSPR is the quality of the database used to develop models. Sorption values in the database were obtained following the experimental procedure described above and are hereafter expressed in mg·g^{-1}. Because measurements have been performed by the same operator on a unique device, it ensures optimal repeatability conditions, leading to minimally

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

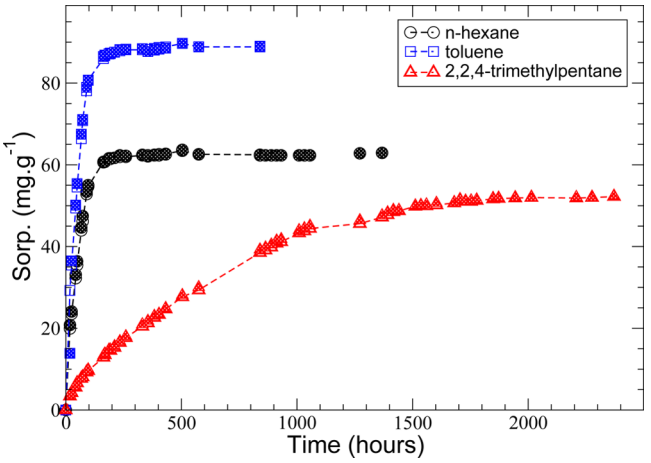


Figure 2. Time evolution of the sorption in a medium density poly(ethylene) for *n*-hexane, toluene, and 2,2,4-trimethylpentane. Empty and filled symbols denote the first and second series of performed sorption measurements, respectively.

noisy experimental values, noting that recent works showed that the robustness of QSPR based models do not necessarily improve when trained upon experimental data from standardized conditions.[35] The database contains 80 sorption values measured at room temperature for neat compounds and up to quinary mixtures of hydrocarbons, alcohols, and ethers in a semicrystalline poly(ethylene). Data points are as follows: 27 neat compounds ranging from C5 to C10, 46 binary mixtures, 5 ternary mixtures, and 2 quinary mixtures. Table 1 presents an extract from this database and gives sorption values measured for neat compounds.

Table 1. Extract from the Database Used in This Work[a]

compound	sorp ($mg \cdot g^{-1}$)	compound	sorp ($mg \cdot g^{-1}$)
n-pentane	57.3	*n*-propylbenzene	76.1
n-hexane	59.2	1-pentene	64.8
n-octane	55.7	1-hexene	63.8
n-decane	60.3	1-heptene	62.9
2-methylpentane	60.3	1-octene	62.1
2,2,4-trimethylpentane	53.3	1-nonene	61.1
cyclohexane	93.6	1-decene	60.1
tetraline	90.1	1,5-cyclooctadiene	92.3
benzene	82.8	methanol	0.7
toluene	87.0	ethanol	1.0
m-xylene	90.0	propan-1-ol	1.3
1,3,5-trimethylbenzene	94.7	butan-1-ol	1.6
1,2,4,5-tetramethylbenzene	98.4	2-methylpropan-1-ol	0.6
ethylbenzene	79.7		

[a]Only sorption values measured for neat compounds are reported.

During the past decade of development of QSPR models, the use of external validation has been shown as necessary to ensure its ability to extrapolate to new compounds, that is, out of the database used for the model development.[36] Its popular version is the *n*-fold cross-validation (*n*-CV) in which the data set is randomly split on approximately equal *n* portions. An aggregate of $(n - 1)$ portions forms the training set on which the predictive model is built, the remaining portion constituting the test set; no data point belonging to external sets is used to derived models. This procedure is repeated *n* times choosing at each new fold another portion of data as a test set. The subject of external validation for QSPR analysis of mixtures has recently been treated by Muratov et al.,[37] and the strategy of external validation applied in this study is "mixture out". We used a 5-CV procedure; consequently, the training and test sets represent 80% and 20% of the database, respectively.

Molecular and Mixture Descriptors. From conclusions drawn in previous studies we chose to work with two sets of descriptors.[1] First, functional group count descriptors (FGCD) gather some counts of groups identified as relevant under chemical aspects. Table 2 gives a list of FGCD under consideration in this study, labeled from X1 to X13. The FGCD labeled X13 denotes the number of carbon atoms involved in a ring. As Patil et al. did,[20] we have also computed the molar mass (MM) of neat compounds in the database, this information being used as an additional descriptor. Such a simple representation of compounds has been shown to provide relevant descriptors usable in QSPR procedures.[2] In the case of FGCD containing at least one heteroatom, we considered their exponentiations as new descriptors. For instance, X12 to the power two, three, and four have been added in the descriptor set. Simplified molecular input line entry specification (SMILES) notations were assigned to each neat compound belonging to the database. FGCD were counted using the Open Babel's SMILES arbitrary target specification (SMARTS) matching functionalities,[38] and SMARTS codes corresponding to FGCD are given in Table 2.

Table 2. List of the Functional Group Count Descriptors (FGCD) Used To Describe Hydrocarbons, Alcohols and Ethers in the Database, And Associated Symbols and SMARTS Codes

symbol	FGCD	SMARTS
C	X1	[C,c]
H	X2	[H]
-CH_3	X3	[CX4H3]
-CH_2-	X4	[CX4H2]
>CH-	X5	[CX4H1]
>C<	X6	[CX4H0]
=CH_2	X7	[C]=[CX3H2]
=CH-	X8	[C]=[CX3H1]
•>CH	X9	[cX3H1]
•>C-	X10	[cX3H0]
C-O-C	X11	[C]-[O]-[C]
OH	X12	[OH]
Nb in ring	X13	[R]
molar_mass	MM	NA

The second set of descriptors was constituted using the ISIDA fragmentor software[39] to identify and count relevant substructural molecular fragments (SMFs).[40,41] To establish SMFs, there exists various types of molecular subgraphs such as sequences and augmented atoms. Sequences consist in successions of atoms and bonds, atoms only, or bonds only in the molecular graph. Augmented atoms stand for a given atom with its nearest neighboring including atoms and bonds, atoms only, bonds only, or atom pairs. These latter represent a kind of extension of FGCD including surrounding of chemical functions. Thus, we generated augmented atoms with fragments containing from two to three atoms and bonds. Table 3 gives a list of SMFs used in this study, labeled from S1 to S40. In the case of SMFs containing at least one heteroatom, we

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

Table 3. List of the Substructural Molecular Fragments (SMF) Used To Describe Hydrocarbons, Alcohols, and Ethers in the Database

SMF	description	SMF	description
S1	(C−C),(C−C−C),xC	S21	(C−C),(C−C),(C−C),(C−C),(C−C−C),xC
S2	(C−C),(C−C),(C−C−C),xC	S22	(C−C),(C−C),(C−C−C),(C−C−C),(C−C−C),(C−C−C),(C−C−C),xC
S3	(C−C),(C−C),(C−C−C),(C−C−C),xC	S23	(C−C),(C−C),(C−C),(C−C−C),(C−C−C),(C−C−C),(C−C−C),xC
S4	(C−C),(C−C═C),(C═C),(C═C−C),xC	S24	(C−C),(C−C−C),(C═C),(C═C−C),xC
S5	(C−C),(C−C),(C−C═C),(C═C),(C═C−C),xC	S25	(C−C),(C−C−O),xC
S6	(C−C),(C−C═C),(C═C),(C═C−C),(C═C−C),xC	S26	(C−C),(C−O),xC
S7	(C−C),(C−C−C),(C−C═C),(C═C),(C═C−C),xC	S27	(O−C),(O−C−C),xO
S8	(C−C),(C−C−C),(C−C═C),xC	S28	(C−O),xC
S9	(C−C),(C−C−C),(C−C═C),(C═C),(C═C−C),(C═C−C),xC	S29	(O−C),xO
S10	(C−C),(C−C),(C−C−C),(C−C═C),(C═C),(C═C−C),xC	S30	(C−C),(C−C),(C−C),(C−C−O),xC
S11	(C−C),(C−C),(C−C═C),(C═C),(C═C−C),(C═C−C),xC	S31	(C−C),(C−C−C),(C−C−C),(C−O),xC
S12	(C−C),(C−C),(C−C−C),(C−C═C),(C═C),(C═C−C),(C═C−C),xC	S32	(C−C),(C−C),(C−C−O),xC
S13	(C−C),(C−C),(C−C−C),(C−C−C),(C−C═C),xC	S33	(C−C),(C−C−C),(C−O),xC
S14	(C−C),(C−C),(C−C−C),(C−C═C),xC	S34	(O−C),(O−C),(O−C−C),(O−C−C),(O−C−C),xO
S15	(C═C),(C═C−C),xC	S35	(C−C),(C−C),(C−C),(C−O),(C−O−C),xC
S16	(C−C),(C−C−C),(C═C),xC	S36	(C−C),(C−C−C),(C−C−C),(C−C−O),xC
S17	(C−C),(C−C−C),(C−C−C),xC	S37	(C−O),(C−O−C),xC
S18	(C−C),(C−C),(C−C),(C−C−C),xC	S38	(C−C),(C−C),(C−C−C),(C−C−O),xC
S19	(C−C),(C−C),(C−C−C),(C−C−C),(C−C−C),xC	S39	(O−C),(O−C),(O−C−C),(O−C−C),(O−C−C),(O−C−C),xO
S20	(C−C),(C−C−C),(C−C−C),(C−C−C),xC	S40	(C−C),(C−O),(C−O−C),xC

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

considered their exponentiations as new descriptors. For instance, S29 to the power two, three, and four have been added in the descriptor set.

To compute FGCD and SMF decriptors and thus extract features of mixtures, we considered in a first approximation only linear combinations of pure component descriptors weighted with the associated molar fractions, x_i. This approach has already been applied with FGCD and SMF for the modeling of mixture properties such as the optimal salinity of surfactants.[42] In the case of descriptor X1, the corresponding descriptor for a mixture $X1_{mix}$, is defined as follows:

$$X1_{mix} = \sum_{i=1}^{N} x_i X1_i \tag{3}$$

where i runs over the N constituants in the mixture. Descriptor values are then standardized using the mean and the standard deviation of the initial descriptor values.

Machine Learning Methods. The genetic function approximation (GFA)[43] as implemented in the Materials Studio software[44] was used to build multilinear models. The GFA was chosen for its ability to identify and combine the most relevant descriptors over a large number of molecular features. The GFA procedure consists in iterations of selections, crossovers, and mutations, coupled with objective criteria such as the well-known coefficient of determination (R^2) in order to extract the best fitting models. In this work, the adjusted R^2 was used as the objective criteria, and k, the maximum number of descriptors that can be included in the final model, was fixed to 14 in order to respect the statistical criteria $n \geq 4k$, where n is the number of data points in the training set.[45] The initial population (i.e., number of equations) was set to 500, and the maximum generation number to 100000. This procedure was performed on each of the five training sets.

The support vector machine (SVM)[46] from the Libsvm package[47,48] was used to generate ϵ-SVM regression models. This method attempts to find a function that fits the data as flatly as possible (by using an ϵ insensitive loss function) while minimizing the number of support vectors; such procedure is known as structural risk minimization. The kernel function $K(x,x')$, used in this work is the radial basis function (RBF) kernel defined as follows:

$$K(x, x') = e^{-\gamma \|x-x'\|^2} \tag{4}$$

where x and x' are sample vectors and γ is a parameter related to the averaged euclidean distance between two instances in the sample. While faster methods have been recently proposed to optimize SVM parameters such as those based on intercluster distances in the feature space,[49−51] the cost (C), γ, and ϵ were optimized using a grid search method. The grid dimensions were chosen so that $C \in [2^{-7}, 2^{-6}, ..., 2^6, 2^7]$, $\gamma \in [2^{-20}, 2^{-19}, ..., 2^{19}, 2^{20}]$, and $\epsilon \in [2^{-8}, ..., 2^8]$; the training data and each $\{C, \gamma, \epsilon\}$ combination are used to train an SVM model. SVM models are then used to classify test data, and the optimal region of the $\{C, \gamma, \epsilon\}$ grid is determined on the basis of both root mean square error (RMSE) and R^2 values. The surrounding of the optimal region of the $\{C, \gamma, \epsilon\}$ grid is then explored in detail, and the best $\{C, \gamma, \epsilon\}$ set of parameters is deduced from RMSE and R^2 values on the test set. For each pool of descriptors, the n-fold cross-validation procedure has been applied twice as recently performed by Muller et al.[42] The general model's performance has been assessed in 5-CV. However, in order to optimize parameters of the SVM method, an additional 6-CV was applied to the training set on each fold.

The bagging technique is then applied, it combines models in such a way that the obtained consensus model is more predictive and robust than individual models, and the consensus model that exhibits the best performances is selected. The selection of best models is performed using statistical criteria such as the average absolute error (AAE), RMSE, R^2, and the concordance correlation coefficient (CCC).[52] Chirico et al. have shown that the use of this latter coefficient is advocated considering various scenarios such as location shifts, scale shifts, and location plus scale shifts.[53,54]

Applicability Domain. Among the numerous existing approaches,[55] the leverage approach (Mahalanobis distance from the structural centroid of data points in the training set) has been followed to verify whether external candidates lay in the applicability domain (AD) of multilinear models.[36] The leverage was measured through h_i, the diagonal elements of the Hat matrix, $\mathbf{H}$, which is defined as follows:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{5}$$

where $\mathbf{X}$ is the matrix of descriptors used in the predictive model for data points belonging to the training set. $\mathbf{H}$ is an $m \times m$ matrix, with m being the number of data points in the training set. For external fluid candidates, Hat values, h_i, were computed using

$$h_i = \chi_i(\mathbf{X}^T\mathbf{X})^{-1}\chi_i^T \tag{6}$$

with χ_i being the vector of descriptors associated with the external fluid candidate i. Williams plots are an easy way to determine whether a fluid candidate lies in the AD[45] and correspond to the plot of the standardized residuals as a function of the Hat values. Using this representation, a prediction is considered as reliable if the Hat value is lower than a limit labeled h^*, defined as follows:

$$h^* = \psi\bar{h} \quad \text{with} \quad \bar{h} = (d+1)/m \tag{7}$$

where ψ is a coefficient used to tune the AD restrictiveness and d and m are the number of descriptors selected in the model and the number of data points in the training set, respectively.

RESULTS AND DISCUSSION

In this section, we report various QSPR models to predict sorption in MDPE a semicrystalline poly(ethylene) for mixtures of hydrocarbons, alcohols, and ethers. Models were derived on each training set and performances of models evaluated on corresponding test sets. Two machine learning methods (i.e., GFA and SVM) and two families of descriptors (i.e., FGCD and SMF) were used, leading to four classes of models labeled as follows: GFA−FGCD, GFA−SMF, SVM−FGCD, and SVM−SMF.

All obtained GFA based models are composed of 13 or 14 variables, which is the upper limit imposed during the development procedure. Equation 8 presents one of the five GFA−FGCD models developed; the meaning of descriptors is given in Table 2.

$$\begin{aligned} \text{sorp} = {} & -758.8X1 + 5.6X7 + 13.6X8 + 37.2X9 \\ & + 20.3X10 - 120.2X12 + 35.9X13 + 606.0MM \\ & - 122.0X11^2 + 85.2X11^3 - 157.6X12^2 \\ & + 264.5X12^3 - 150.9X12^4 + 62.1 \end{aligned} \tag{8}$$

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

Table 4. Performance Characteristics for Models Developed Using GFA and SVM Approaches and FGCD and SMF Sets of Descriptors[a]

set		GFA−FGCD	GFA−SMF	SVM−FGCD	SVM−SMF
training	AAD	3.6(0.2)	3.7(0.1)	3.1(0.9)	2.9(1.4)
	RMSE	4.8(0.2)	5.7(0.1)	3.9(1.2)	5.2(0.7)
	R^2	0.960(0.004)	0.943(0.005)	0.972(0.019)	0.953(0.012)
	CCC	0.980(0.002)	0.970(0.003)	0.985(0.011)	0.976(0.006)
test	AAD	6.5(0.9)	6.1(1.2)	6.0(1.0)	6.0(1.9)
	RMSE	10.0(2.8)	7.9(1.8)	9.3(2.5)	8.1(1.9)
	R^2	0.662(0.111)	0.784(0.077)	0.710(0.081)	0.753(0.175)
	CCC	0.773(0.094)	0.892(0.039)	0.819(0.064)	0.885(0.054)

[a]Performance are indicated for the training (64 data points) and test (16 data points) sets. Values stand for means of performance characteristics obtained on the five training/test splittings, and values in parentheses denote standard deviations.

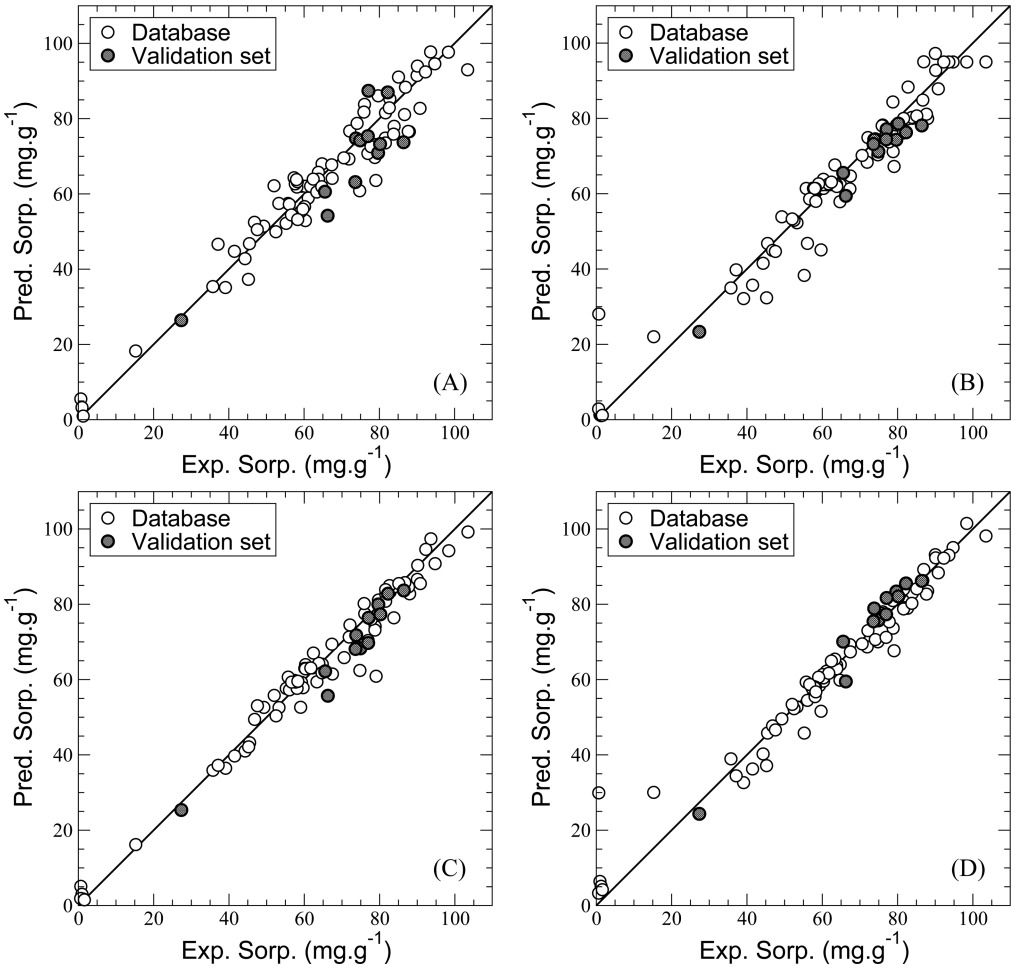


Figure 3. Scatterplots of experimental sorptions vs predicted sorptions using GFA−FGCD (A), GFA−SMF (B), SVM−FGCD (C), and SVM−SMF (D) models. Database stands for the 80 fluid candidates, and the validation set contains complex mixtures representative of gasoline fuels.

Equation 8 is in line with some physical intuitions. For instance, increment of the number of carbon atoms (X1) results in lower sorption values, unsaturation (X7 and X8) increases sorption values, and a naphthenic or an aromatic ring (X13) contributes to higher sorption values. Table 4 presents for developed models averages of performance characteristics computed over the five training/test splittings. GFA−FGCD and GFA−SMF models show similar performances after the training stage. Standard deviations reported in Table 4 indicate that performances for these two classes of models do not vary significantly with training/test splittings. SVM−FGCD and SVM−SMF show very similar performances after the training set stage. From standard deviation values, it seems that performances of at least one of splittings lead to overall performance losses. When applied to external sets the SVM based models seemed to roughly perform better than GFA−FGCD and GFA−SMF models. In the case of external sets, RMSE standard deviations computed for GFA based models are much more important than in the case of the training set.

For each of the four classes of models, a consensus modeling approach was followed to combine predictions of models obtained from the five training/test splittings. Figure 3A−D presents scatterplots of experimental sorptions vs predicted sorptions for sorption values in the database (white circles),

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

Table 5. Performance Characteristics for the Consensus Models' Predictions[a]

set		GFA–FGCD	GFA–SMF	SVM–FGCD	SVM–SMF
database	AAD	4.1	4.0	2.8	2.9
	RMSE	5.4	5.8	3.9	4.9
	R^2	0.945	0.936	0.971	0.953
	CCC	0.972	0.967	0.985	0.975
validation	AAD	6.5	3.3	3.7	3.0
	RMSE	7.9	4.3	4.8	3.6
	R^2	0.706	0.914	0.893	0.940
	CCC	0.869	0.958	0.950	0.973

[a]Database stands for the 80 fluid candidates, and the Validation set contains complex mixtures representative of gasoline fuels.

and Table 5 presents performance characteristics for best consensus models computed considering all sorption values in the database. Clearly, the consensus model SVM−FGCD performs better than the GFA−FGCD, and the SVM−SMF performs better than the GFA−SMF on the database (i.e., the 80 data points). No huge difference appears between FGCD and SMF based models for predictions on the database, except a data point with a very low sorption value for which the used pool of SMF descriptors lead to an overestimation of the experimental sorption value.

Figure 4A,B presents Williams plots for GFA−FGCD and GFA−SMF models, respectively. Following this approach, only a few data points are considered as out of the applicability domains of these two models. In the case of the GFA−FGCD model, three neat compounds (1,2,4,5-tetramethylbenzene, 2,2,4-trimethylbenzene, and cyclooacta-1,5-diene) and one mixture (EtBE/EtOH, 0.31:0.69 mol %) are out of the AD. One can remark that these predicted sorption values are in good agreement with corresponding experimental data. In the case of the GFA−SMF model, three neat compounds (benzene, 2,2,4-trimethylbenzene, and ethanol) and one mixture (ETBE/EtOH, 0.31:0.69 mol %) are out of the AD. The predicted sorption values for these data points are in good agreement with corresponding experimental data. It is not surprising that some of the neat compounds appear as out of the AD because these data points are located at the outer part of the database chemical space. This can be graphically observed using a projection of data points on the two first principal components from a principal component analysis. In Figure 4A, all standardized residual of data points are in absolute lower than 3σ. Figure 4B indicates that the GFA−SMF failed to predict the sorption of methanol with an absolute standardized residual greater than 4σ.

An additional external validation of models was performed measuring sorptions of 12 gasolines in the semicrystalline poly(ethylene). The studied gasoline are based on three commercial gasolines in blends with various amounts (up to 85%) of ethanol, MTBE, and ETBE. The compositions of the three commercial gasolines were determined from the post-treatement of gas chromatographic analysis using the Carburane software, resulting in three complex mixtures containing *circa* 300 compounds. FGCD and SMF were computed for each of the 300 neat compounds, and FGCD and SMF for the 12 complex mixtures were computed using eq 3. Williams plots presented in Figure 4A,B indicate that gasolines lie in the AD of models. Projections of data points on the two first principal components of a principal component

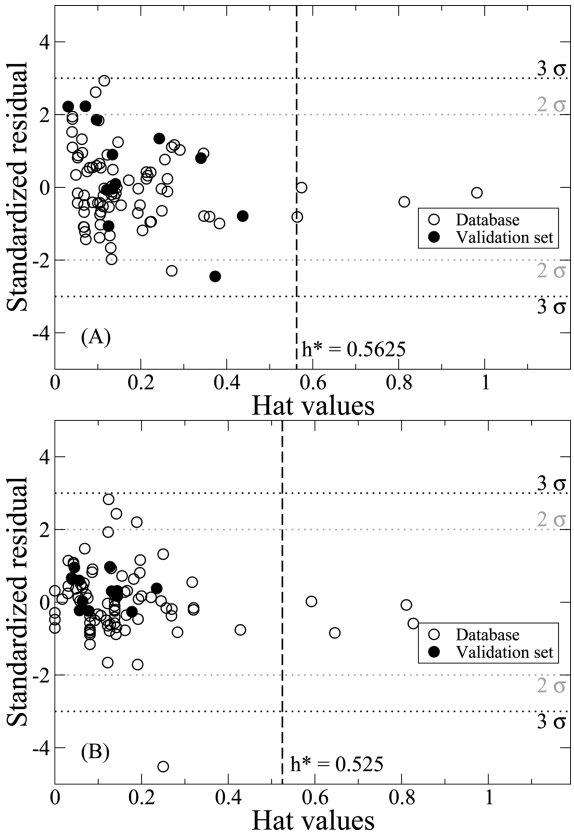


Figure 4. Williams plot for GFA−FGCD (A) and GFA−SMF (B) models. Database stands for the 80 fluid candidates, and the validation set contains complex mixtures representative of gasoline fuels. The cutoff limits at 2σ and 3σ (with the standard deviation, σ, equal to 1) are indicated as horizontal dotted lines, and the critical hat value h^*(for ψ = 3, see eq 7) is represented with the vertical dashed line.

analysis revealed that gasolines lie in the center of the chemical space. Figure 3A−D presents scatterplots of experimental sorptions vs predicted sorptions for the 12 gasolines (filled circles), and Table 5 presents performance characteristics for best consensus models computed considering sorption values for studied gasolines (validation set). None of the four predictive models failed in the prediction of the sorption values of gasolines. Among the four models, the one leading to the best predictions is clearly the SVM−SMF model. Although the used pool of SMF descriptors failed in the description of light alcohols, the SMF based models succeed in the prediction of the sorption (27 mg·g^{-1}) for a gasoline in mixture with 85 vol % of ethanol.

The SVM−SMF model represents a useful tool to quickly estimate for a new compound or mixture its sorption in a semicrystalline poly(ethylene) and thus rapidly check materials compatibility, avoiding thousands hours of experiments. For instance, a synthetic fuel candidate molecule, 2,6,10-trimethyldodecane (farnesane), is currently under consideration as an alternative jet fuel and diesel.[56] The use of our SVM−SMF model returns an estimated sorption of 55 mg·g^{-1}.

The SVM−SMF model can be used to draw tendencies of the sorption evolution when increasing the number of carbon atoms in some hydrocarbons and *n*-alkan-1-ol (see Figure 5). For all studied families of hydrocarbons, the SVM−SMF model returns a decrease of the sorption when the number of carbon atoms inceases. A comparison of the branching effect shows

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

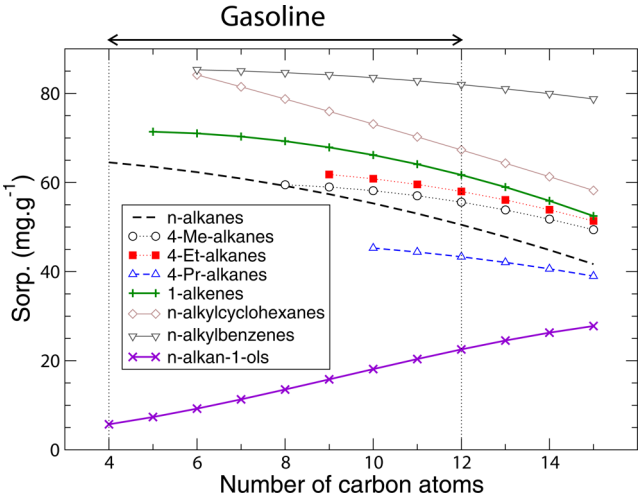


Figure 5. Tendencies of the sorption evolution for some hydrocarbons and *n*-alkan-1-ol when increasing the number of carbon atoms. Due the slight spreading in predictions, data have been fitted using polynomial equations. Gasoline stands for the range of carbon atom numbers in such fluids.

that at the same number of carbon atoms, 4-Me-alkanes and 4-Et-alkanes have similar sorption values. However, when the size of the alkyl group increases, Figure 5 shows that the sorption of 4-Pr-alkanes is ca. 20% lower than that of 4-Me-alkanes and 4-Et-alkanes. Figure 5 proposes a comparison between the evolution of sorption for linear hydrocarbons showing that the sorption of 1-alkenes is about 20% greater than that of its saturated counterpart. Figure 5 also shows that while the sorption of *n*-alkylbenzenes is roughly constant for up to C12 compounds, a linear decrease is predicted for *n*-alkylcyclohexanes. For *n*-alkan-1-ol molecules, the sorption in a semicrystalline poly(ethylene) is reported to increase with the number of carbon atoms, noting that sorption values for long chain *n*-alkan-1-ols tend to match with those of *n*-alkanes. Randová et al. have recently reported similar observations regarding the sorption of pure liquids into polymeric membrane.[25]

Figure 6 shows predictions carried out using the SVM−SMF model for methanol/toluene binary mixtures. As discussed previously, the sorption of methanol is not well predicted. Predicted values presented on Figure 6 are in a relatively good

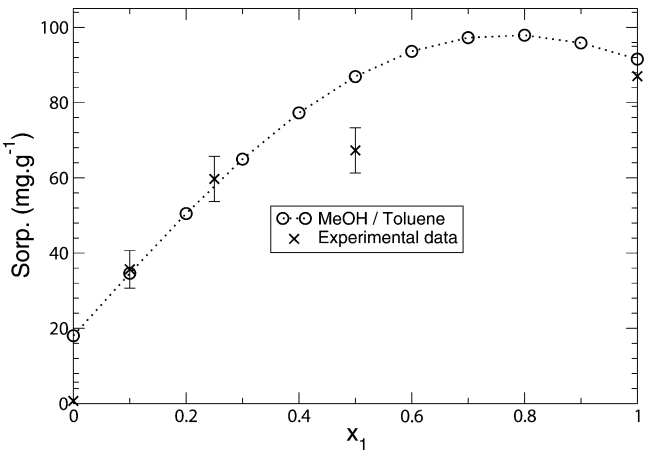


Figure 6. Prediction of sorption variations for methanol−toluene binary mixtures, where x_1 is the toluene mole fraction.

agreement with most corresponding experimental data points. The trend of our experimental data is in line with that observed by Randová et al.[23,24] A maximum sorption behavior is predicted by the SVM-SMF model for the methanol/toluene binary mixture, and the predicted maximum sorption (86 mg·g^{-1}) is reached at *circa* 0.8 mole fraction of toluene. For mole fraction of toluene values greater than 0.4, our model seems to overestimate the sorption of the fluid in a semicrystalline poly(ethylene).

CONCLUSION

Machine learning approaches have been used to model the sorption of neat compounds and up to quinary mixtures of some hydrocarbons, alcohols, and ethers in a semicrystalline poly(ethylene). Experimental data were obtained using our original experimental apparatus. The generated database has been analyzed with chemoinformatics tools, and combinations of two machine learning methods (i.e., GFA and SVM) and two families of descriptors (i.e., FGCD and SMF) were used to derive predictive models.

In addition to the usual internal/external validations, the model was further tested using new experimental data for gasolines, and the predictions were in excellent agreement with the data. We performed some predictions for the sorption variations upon increase of the number of carbon atoms in a series of hydrocarbons and for *n*-alkan-1-ols. We also studied methanol−toluene binary mixtures containing various amounts of toluene and observed a maximum sorption behavior.

Polymers are widely used in various industrial applications such as packaging, car industry, and membrane separation, among others.[57] The determination of the sorption of gases and liquids in polymers is fundamental and must be known before any applications. Our work shows that when a good quality database and various machine learning approaches are used and consensus modeling is applied, the so-obtained predictive models are powerful tools to estimate a property, in this case the sorption of chemicals in a semicrystalline poly(ethylene). The combinatorial use of experiments and chemoinformatics tools contributes to drastically reducing the time necessary to quantify polymeric materials compatibility with a fluid candidate according to its structural characteristics. This work is a solid step in the efforts of *in silico* determination of fluid sorption in polymers and is to be extended to various families of polymers, conditions of temperature and pressure, and larger ranges of carbon atom numbers for penetrant chemicals.

AUTHOR INFORMATION

Corresponding Author

*E-mail: benoit.creton@ifpen.fr.

Present Address

‡N.V.: AXENS, 89 boulevard Franklin Roosevelt, 92508 Rueil-Malmaison, France.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully thank Drs. Laurie Starck and Christophe Muller for helpful discussions.

REFERENCES

(1) Saldana, D. A.; Creton, B.; Mougin, P.; Jeuland, N.; Rousseau, B.; Starck, L. Rational formulation of alternative fuels using QSPR

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

methods: Application to jet fuels. *Oil Gas Sci. Technol.* **2013**, *68*, 651−662.
(2) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Pidol, L.; Jeuland, N.; Creton, B. Flash point and cetane number predictions for fuel compounds using quantitative structure property relationship (QSPR) methods. *Energy Fuels* **2011**, *25*, 3900−3908.
(3) Surisetty, V. R.; Dalai, A. K.; Kozinski, J. Alcohols as alternative fuels: An overview. *Appl. Catal., A* **2011**, *404*, 1−11.
(4) Maru, M. M.; Lucchese, M. M.; Legnani, C.; Quirino, W. G.; Balbo, A.; Aranha, I. B.; Costa, L. T.; Vilani, C.; de Sena, L. Á; Damasceno, J. C.; dos Santos Cruz, T.; Lidízio, L. R.; Ferreira e Silva, R.; Jorio, A.; Achete, C. A. Biodiesel compatibility with carbon steel and HDPE parts. *Fuel Process. Technol.* **2009**, *90*, 1175−1182.
(5) Zagidullin, R. N.; Idrisova, V. A.; Dmitrieva, T. G.; Gil'Mutdinov, A. T. Additive package for alternative automotive fuels. *Chem. Technol. Fuels Oils* **2011**, *47*, 183−187.
(6) Subramanian, P. M. In *Barrier Polymers and Structures*; Koros, W. J., Ed.; ACS Symposium Series; American Chemical Society: Washington, DC, 1990; Vol. *423*, Chapter 13, pp 252−265.
(7) Berlanga-Labari, C.; Albistur-Goñi, A.; Barado-Pardo, I.; Gutierrez-Peinado, M.; Fernández-Carrasquilla, J. Compatibility study of high density polyethylene with bioethanol-gasoline blends. *Mater. Eng.* **2011**, *32*, 441−446.
(8) Kass, M. D.; Janke, C.; Theiss, T.; Pawel, S.; Baustian, J.; Wolf, L.; Koch, W. Compatibility Assessment of Plastic Infrastructure Materials to Test Fuels Representing Gasoline Blends Containing Ethanol and Isobutanol. *SAE International Journal of Fuels and Lubricants* **2014**, *7*, 457−470.
(9) Kallio, K. J.; Nageye, A. S.; Hedenqvist, M. S. Ageing properties of car fuel-lines; accelerated testing in "close-to-real" service conditions. *Polym. Test.* **2010**, *29*, 41−48.
(10) Yeh, J.-T.; Chen, H.-Y.; Tsai, F.-C. Gasoline permeation resistance of polypropylene, polypropylene/ethylene vinyl alcohol, polypropylene/modified polyamide, and polypropylene/blends of modified polyamide and ethylene vinyl alcohol containers. *J. Polym. Res.* **2006**, *13*, 451−460.
(11) Gagnard, C.; Germain, Y.; Keraudren, P.; Barrière, B. Permeability of semicrystalline polymers to toluene/methanol mixture. *J. Appl. Polym. Sci.* **2003**, *90*, 2727−2733.
(12) Sato, Y.; Fujiwara, K.; Takikawa, T.; Sumarno; Takishima, S.; Masuoka, H. Solubilities and diffusion coefficients of carbon dioxide and nitrogen in polypropylene, high-density polyethylene, and polystyrene under high pressures and temperatures. *Fluid Phase Equilib.* **1999**, *162*, 261−276.
(13) Faure, F.; Rousseau, B.; Lachet, V.; Ungerer, P. Molecular simulation of the solubility and diffusion of carbon dioxide and hydrogen sulfide in polyethylene melts. *Fluid Phase Equilib.* **2007**, *261*, 168−175.
(14) Memari, P.; Lachet, V.; Rousseau, B. Molecular simulations of the solubility of gases in polyethylene below its melting temperature. *Polymer* **2010**, *51*, 4978−4984.
(15) Memari, P.; Lachet, V.; Klopffer, M.-H.; Flaconnèche, B.; Rousseau, B. Gas mixture solubilities in polyethylene below its melting temperature: Experimental and molecular simulation studies. *J. Membr. Sci.* **2012**, *390*−*391*, 194−200.
(16) Shah, M. R.; Yadav, G. D. Prediction of sorption in polymers using quantum chemical calculations: Application to polymer membranes. *J. Membr. Sci.* **2013**, *427*, 108−117.
(17) Memari, P.; Lachet, V.; Rousseau, B. Gas Permeation in semicrystalline polyethylene as studied by molecular simulation and elastic model. *Oil Gas Sci. Technol.* **2015**, *70*, 227−235.
(18) Creton, B.; Nieto-Draghi, C.; Pannacci, N. Prediction of surfactants' properties using multiscale molecular modeling tools: A review. *Oil Gas Sci. Technol.* **2012**, *67*, 969−982.
(19) Teplyakov, V.; Meares, P. Correlation aspects of the selective gas permeabilities of polymeric materials and membranes. *Gas Sep. Purif.* **1990**, *4*, 66−74.
(20) Patil, G. S.; Bora, M.; Dutta, N. N. Empirical correlations for prediction of permeability of gases/liquids through polymers. *J. Membr. Sci.* **1995**, *101*, 145−152.
(21) Izák, P.; Bartovská, L.; Friess, K.; Šípek, M.; Uchytil, P. Comparison of various models for transport of binary mixtures through dense polymer membrane. *Polymer* **2003**, *44*, 2679−2687.
(22) Izák, P.; Bartovská, L.; Friess, K.; Šípek, M.; Uchytil, P. Description of binary liquid mixtures transport through non-porous membrane by modified Maxwell-Stefan equations. *J. Membr. Sci.* **2003**, *214*, 293−309.
(23) Randová, A.; Bartovská, L.; Hovorka, Š; Friess, K.; Izák, P. The membranes (Nafion and LDPE) in binary liquid mixtures benzene +methanol - sorption and swelling. *Eur. Polym. J.* **2009**, *45*, 2895−2901.
(24) Randová, A.; Bartovská, L.; Hovorka, Š; Izák, P.; Friess, K.; Janku, J. Sorption of binary mixtures of toluene + lower aliphatic alcohols C1 - C6 in low-density polyethylene. *J. Appl. Polym. Sci.* **2011**, *119*, 1781−1787.
(25) Randová, A.; Bartovská, L.; Friess, K.; Hovorka, Š; Izák, P. Fundamental study of sorption of pure liquids and liquid mixtures into polymeric membrane. *Eur. Polym. J.* **2014**, *61*, 64−71.
(26) Randová, A.; Bartovská, L.; Izák, P.; Friess, K. A new prediction method for organic liquids sorption into polymers. *J. Membr. Sci.* **2015**, *475*, 545−551.
(27) Creton, B.; Dartiguelongue, C.; De Bruin, T.; Toulhoat, H. Prediction of the cetane number of diesel compounds using the quantitative structure property relationship. *Energy Fuels* **2010**, *24*, 5396−5403.
(28) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Creton, B. Prediction of flash points for fuel mixtures using machine learning and a novel equation. *Energy Fuels* **2013**, *27*, 3811−3820.
(29) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Ferrando, N.; Creton, B. Prediction of density and viscosity of biofuel compounds using machine learning methods. *Energy Fuels* **2012**, *26*, 2416−2426.
(30) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Creton, B. On the rational formulation of alternative fuels: melting point and net heat of combustion predictions for fuel compounds using machine learning methods. *SAR and QSAR in environmental research* **2013**, *24*, 259−277.
(31) The Carburane software is a IFP Energies nouvelles product. Available online http://ifp-energies-nouvelles1.software.informer.com/.
(32) Durand, J. P.; Fafet, A.; Barreau, A. Direct and automatic capillary GC analysis for molecular weight determination and distribution in crude oils and condensates up to C20. *J. High Resolut. Chromatogr.* **1989**, *12*, 230−233.
(33) Durand, J. P.; Beboulene, J. J.; Ducrozet, A.; Bre, A.; Carbonneaux, S. Improvement of Simulated Distillation Methods by Gas Chromatography in Routine Analysis. *Oil Gas Sci. Technol.* **1999**, *54*, 431−438.
(34) Nieto-Draghi, C.; Bocahut, A.; Creton, B.; Have, P.; Ghoufi, A.; Wender, A.; Boutin, A.; Rousseau, B.; Normand, L. Optimisation of the dynamical behaviour of the anisotropic united atom model of branched alkanes: application to the molecular simulation of fuel gasoline. *Mol. Simul.* **2008**, *34*, 211−230.
(35) Palmer, D. S.; Mitchell, J. B. O. Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? *Mol. Pharmaceutics* **2014**, *11*, 2962−2972.
(36) Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694−701.
(37) Muratov, E. N.; Varlamova, E. V.; Artemenko, A. G.; Polishchuk, P. G.; Kuz'Min, V. E. Existing and developing approaches for QSAR analysis of mixtures. *Mol. Inf.* **2012**, *31*, 202−221.
(38) SMARTS - A Language for Describing Molecular Patterns; Daylight Chemical Information Systems Inc.: Laguna Niguel, CA.; http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, accessed in 2014.
(39) http://infochim.u-strasbg.fr, accessed in 2014.

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640

(40) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: An universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693−703.
(41) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-labelled fragment descriptors. *Mol. Inf.* **2010**, *29*, 855−868.
(42) Muller, C.; Maldonado, A. G.; Varnek, A.; Creton, B. Prediction of optimal salinities for surfactant formulations using a Quantitative Structure-Property Relationships approach. *Energy Fuels* **2015**, *29*, 4281−4288.
(43) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Model.* **1994**, *34*, 854−866.
(44) Materials Studio. version 5.5, Accelrys Software Inc.: San Diego, USA, 2014.
(45) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(46) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*; Cambridge University Press: New York, NY, USA, 2000.
(47) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2011**, *2*, 27.
(48) LIBSVM website. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, accessed in 2014.
(49) Wu, K.-P.; Wang, S.-D. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. *Pattern Recognition* **2009**, *42*, 710−717.
(50) Zhang, X.; Zhou, J.; Wang, C.; Li, C.; Song, L. Multi-class support vector machine optimized by inter-cluster distance and self-adaptive deferential evolution. *Applied Mathematics and Computation* **2012**, *218*, 4973−4987.
(51) Zhang, X.; Qiu, D.; Chen, F. Support vector machine with parameter optimization by a novel hybrid method and its application to fault diagnosis. *Neurocomputing* **2015**, *149* (Part B), 641−651.
(52) Lin, L. I.-K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, *45*, 255−268.
(53) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: How to evaluate It? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320−2335.
(54) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* **2012**, *52*, 2044−2058.
(55) Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of {QSAR} models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22−29.
(56) Won, S. H.; Dooley, S.; Veloo, P. S.; Wang, H.; Oehlschlaeger, M. A.; Dryer, F. L.; Ju, Y. The combustion properties of 2,6,10-trimethyl dodecane and a chemical functional group analysis. *Combust. Flame* **2014**, *161*, 826−834.
(57) Kimmerlin, G. Interactions fluids polymers: Permeability, durability. *Oil Gas Sci. Technol.* **2015**, *70*, 219−225.

DOI: 10.1021/acscombsci.5b00094
ACS Comb. Sci. 2015, 17, 631−640